Tail probabilities of the delay in a batch-service queueing model with batch-size dependent service times and a timer mechanism

Dieter Claeys, Bart Steyaert, Joris Walraevens[1], Koenraad Laevens, Herwig Bruneel

Stochastic Modelling and Analysis of Communication Systems (SMACS) Research Group, Department of Telecommunications and Information Processing (TELIN), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
Tel.: +32 9 264 3411
Fax: +32 9 264 4295

Abstract

We deduce approximations for the tail probabilities of the customer delay in a discrete-time queueing model with batch arrivals and batch service. As in telecommunications systems transmission times are dependent on packet sizes, we consider a general dependency between the service time of a batch and the number of customers within it. The model also incorporates a timer mechanism to avoid excessive delays stemming from the requirement that a service can only be initiated when the number of present customers reaches or exceeds a service threshold. The service discipline is first-come, first-served (FCFS). We demonstrate in detail that our approximations are very useful for the purpose of assessing the order of magnitude of the tail probabilities of the customer delay, except in some special cases that we discuss extensively. We also illustrate that neglecting batch-size dependent service times or a timer mechanism can lead to a devastating assessment of the tail probabilities of the customer delay, which highlights the necessity to include these features in the model. The results from this paper can, for instance, be applied to assess the quality of service (QoS) of Voice over IP (VoIP) conversations, which is typically expressed in terms of the order of magnitude of the probability of

Email address: Dieter.Claeys@telin.ugent.be (Dieter Claeys)

[1]The third author is a Postdoctoral Fellow with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

packet loss due to excessive delays.

Keywords: batch service, batch arrivals, batch-size dependent, timer, customer delay, tail probabilities

---

1. Introduction

In many real-life circumstances, customers receive some kind of service in group, which is often referred to as batch service. An elevator can be conceived as a textbook example, since elevators can convey several people simultaneously to another floor. Other examples include transport vehicles, busses, ship locks, ovens in production processes, attractions in amusement parks, et cetera. Furthermore, in telecommunications, it is often the case that information packets are grouped in larger entities (batches) and these batches are transmitted instead of each packet individually. This is mainly done for efficiency reasons, since only one header per aggregated batch has to be constructed, instead of one header per single information unit, thus leading to an increased throughput. Technologies using packet aggregation include Optical Burst Switched (OBS) networks [1], [2] and IEEE 802.11n wireless local area networks (WLANs) [3]. More applications can, for instance, be found in [4].

On account of the wide area of applications, queueing models with batch service have attracted considerable attention. However, the focus was mainly put on the number of waiting customers (see e.g. [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]), whereas the waiting time of customers, also called customer delay, has attracted very few attention, especially in the case of batch arrivals.
In [17], [18] and [19] we have computed the probability generating function (PGF) of the customer delay in distinct discrete-time queueing models with batch arrivals and batch service. Although the established PGFs allow us to calculate various moments of the customer delay, these are not suitable to extract tail probabilities. Nevertheless, this is an important performance measure. For instance, the quality of service (QoS) of Voice Over IP (VoIP) conversations is generally expressed in terms of the (order of magnitude of the) probability that packets arrive too late at the end user (see e.g. [20]). The tail probabilities of the delay in a batch-service queueing model can, among others, be applied to assess the QoS of VoIP conversations in wireless

personal area networks (WPANs). The queueing model then represents a node's output and transmission buffer corresponding to a particular destination and QoS: the output buffer is the queue of the batch-service queueing model, the transmission buffer is the (batch) server (one typically places bursts instead of individual packets in the transmission buffer to increase the throughput), and the time that a burst resides in the transmission buffer is the service time.

In view of this, we have established in [21] an approximation for the tail probabilities of the customer delay in a batch-arrival, batch-service queueing model with single-slot service times and with a server that only serves full batches (i.e., the server only starts service when at least as many customers are present as the server capacity). In [22], we have considered a more versatile model with a minimum batch size (also called service threshold) l (i.e., service is initiated only if at least l customers are present, with l some value between 1 and the server capacity) and generally distributed service times. In this paper, we extend our previous work [22]. In [22], the service times do not depend on the batch sizes, whereas in actual telecommunications systems transmission times depend on packet sizes. In addition, it has been shown in [22] that in case of light traffic, the delay can be extremely high when a minimum batch size is enforced. Therefore, in the model studied in this paper, we consider a *general dependency between the service time of a batch and the number of customers within it*, and we include a *timer mechanism* that avoids excessive delays in case of light traffic as well. It will turn out that the analysis of these extensions entail various pitfalls and that neglecting those pitfalls leads to inaccurate approximations. In addition, we focus more on an extensive evaluation of the accuracy of our approach. We demonstrate that the established approximations are very useful to assess the order of magnitude of the tail probabilities of the customer delay, except in some peculiar situations which we discuss in detail. Finally, we illustrate that neglecting batch-size dependent service times or a timer mechanism can lead to distorted results, which reflects the importance of including these features in the model.

The remainder of the paper is structured as follows: in section 2 we describe the model. Then, in section 3, we deduce approximations for the tail probabilities. The accuracy of our approach is evaluated extensively in section 4 and the importance of the model is discussed in section 5. Finally, we draw some conclusions in section 6.

2. Model description

We consider a discrete-time queueing model. As such, the time axis is divided into fixed-length contiguous time periods, called slots. Customer arrivals during consecutive slots are modelled by a sequence of independent and identically distributed (IID) random variables, with common random variable A whose probability generating function (PGF) is denoted by $A(z)$. The mean value, often referred to as mean arrival rate, is characterized by λ and is by definition equal to $A^{'}(1)$ (we use primes to indicate derivatives). Customers queue up in awaitance of service in a queue of infinite size. The server can serve batches containing up to c customers. We refer to c as the server capacity. Whenever the server is available at the beginning of a slot and finds less than l customers ($l \leq c$), service is initiated with probability β and postponed with probability $1 - \beta$. If, on the other hand, at least l customers are present, a service is initiated of a batch containing a maximum of c customers. Service times are synchronized with respect to the slot boundaries, i.e., services always start and end at slot boundaries. Hence, service times last an integral number of slots. The service time of a batch containing n customers is represented by T_n and its corresponding PGF by $T_n(z)$. Under these assumptions, $T_0(z)$ describes the length of a server interruption in an empty system. Finally, the service discipline is first-come, first-served (FCFS).

The results in this paper are valid under the following assumptions:

Assumption 1. *The load* $\rho \triangleq \lambda T_c^{'}(1)/c < 1$.

This ensures stability of the system.

Assumption 2. *The radius of convergence of each PGF is strictly larger than 1.*

This implies that all order moments are finite and can be calculated by means of the moment generating property of PGFs. We designate the radius of convergence of some random variable X by $\Re_X$. In addition, we define $\Re_n$ as the radius of convergence of $T_n(A(z))$ and $\Re \triangleq \min\{\Re_n : 0 \leq n \leq c\}$ and $\Re_T \triangleq \min\{\Re_{T_n} : 0 \leq n \leq c\}$.

Assumption 3. $\Re_n \leq \Re_A$, $n = 0, \ldots, c$.

It is worth mentioning that we believe that this assumption is actually a fact, as we have not been able to construct one counterexample[2]. However, as it is tedious to prove that $\Re_n \leq \Re_A$, we mention it as an assumption.

Assumption 4. *$z^c - T_c(A(z))$ is aperiodic, i.e., the highest common factor of the set of integers $\left\{\{c\} \cup \left\{n \in \mathbb{N} : \frac{d^n}{dz^n} T_c(A(z))\Big|_{z=0} \neq 0\right\}\right\}$ equals 1.*

This assumption ensures that the c unknown boundary probabilities $d(n)$, $n = 0, \ldots, c-1$ (see further) are solutions of a set of c linear independent equations. We thus exclude some special cases (for instance when $c = 2k$, $l = c$, $\beta = 0$ and $A(z) = \sum_{n=0}^{\infty} \Pr[A = 2n] z^{2n}$) in order to present a general solution technique.

Assumption 5. $\lim_{z \uparrow \Re} T_c(A(z))/z^c > 1$.

This assumption will assure that $z^c - T_c(A(z))$ has a zero in the interval $]1, \Re[$. We will show that this entails that the tail probabilities of the customer delay are not dominated by a specific dominant singularity of $T_c(A(z))$ (if any). Although we thus exclude some PGFs $T_c(A(z))$, the commonly adopted PGFs satisfy this assumption. The main advantage is that we can present a general solution whereas otherwise an ad hoc approach would have to be adopted for each PGF $T_c(A(z))$.

3. Deduction of approximation formulas

The delay of a randomly tagged customer is defined as the length of the time period, starting at the end of the slot of arrival, until the customer's batch starts receiving service. It can thus be expressed as an integral number of slots.
In [22] for a system without timer mechanism, we have decomposed the delay W of a randomly tagged customer as the maximum of two parts:

$$W = \max(W_1, W_2) \quad .$$

The *queueing delay* W_1 is the time, starting at the beginning of the slot following the slot wherein the tagged customer arrives (i.e., at the same instant

[2]When trying to construct a counterexample, one should verify that the constructed $A(z)$ and $T_n(z)$ are actually PGFs, by checking the normalization condition and verifying that the coefficients in the Taylor series expansions of $A(z)$ and $T_n(z)$ about $z = 0$ are probabilities.

that W starts), to serve batches of customers that have arrived before the tagged customer. The *postponing delay* W_2 is the time, starting at the same moment as the queueing delay, until the batch with the tagged customer contains at least l customers. In this particular case, the actual service of a customer can start only if all preceding batches have been processed (FCFS) and if its own batch contains at least l customers; hence the equation $W = \max(W_1, W_2)$.
It seems natural to follow a similar approach, by simply redefining W_2 somewhat to include the timer mechanism. The postponing delay would then represent the time, starting at the same moment as the queueing delay, until the batch containing the tagged customer contains at least l customers or until the timer has expired. As a result, similar approximations as in [22] would be obtained: the lower bound

$$\Pr[W > w] \geq \max(\Pr[W_1 > w], \Pr[W_2 > w]) \ ,$$

and the upper bound

$$\Pr[W > w] \leq \Pr[W_1 > w] + \Pr[W_2 > w] \ .$$

This approach, however, is incorrect. The reason is that the timer mechanism only *runs after* the queueing delay. Indeed, the timer is started only when the server becomes available and finds less than l customers, whereas when we include the timer in the postponing delay, we implicitly assume that the timer is already counting *during* the queueing delay. Therefore, we have to resort to another approach.

The idea is to *disconnect the postponing delay from the timer mechanism.* We therefore let $\hat{W}_2$ represent the number of slots until the batch containing the tagged customer can be filled with at least l customers. Next, define Θ as the time period, starting immediately *after* the queueing delay W_1, until the timer has expired. The random variable Θ is by definition geometrically distributed, with probability distribution

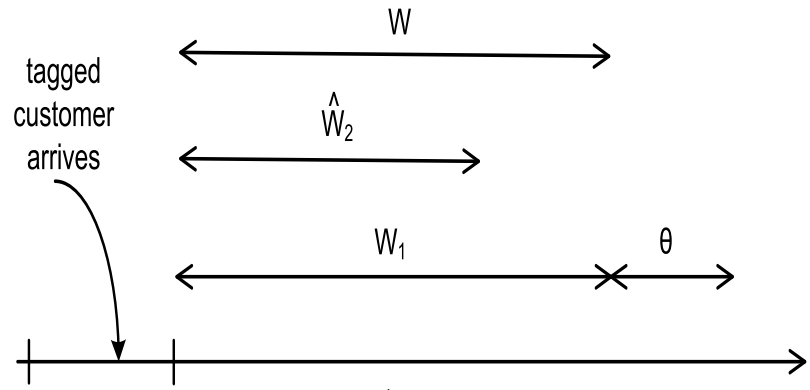
$$\Pr[\Theta = n] = \beta(1-\beta)^n \ , \quad n \geq 0 \ ,$$

or, equivalently,

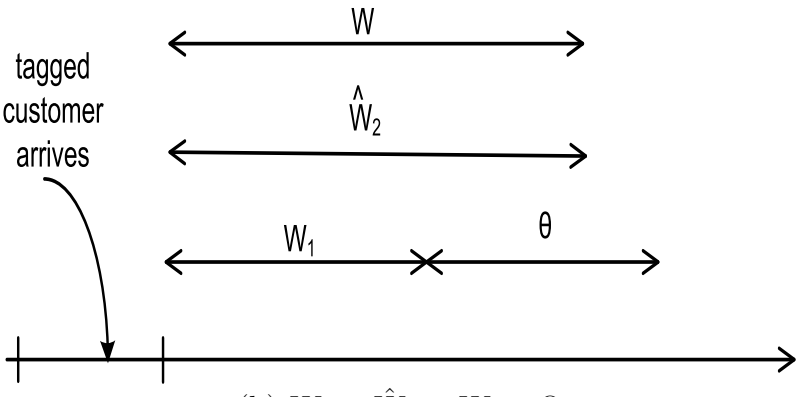
$$\Pr[\Theta > n] = (1-\beta)^{n+1} \ , \quad n \geq 0 \ .$$

On account of these definitions, we have the following relation between W, W_1, $\hat{W}_2$ and Θ (see also Fig. 1):

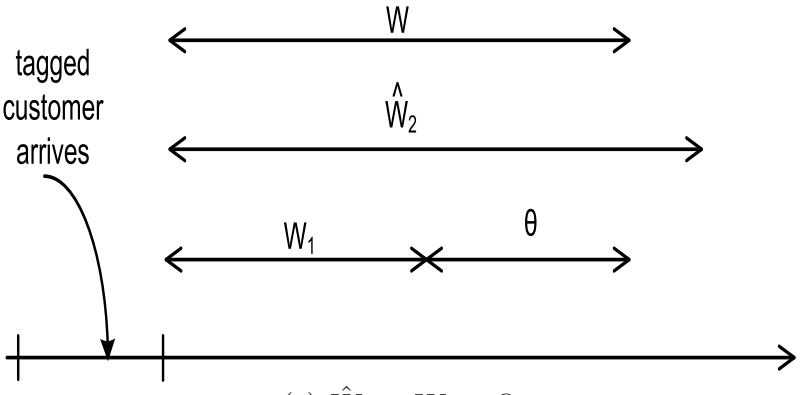
$$W = \begin{cases} W_1 & \text{if } \hat{W}_2 \leq W_1 \ , \\ \hat{W}_2 & \text{if } W_1 < \hat{W}_2 \leq W_1 + \Theta \ , \\ W_1 + \Theta & \text{if } \hat{W}_2 > W_1 + \Theta \ . \end{cases}$$



(a) $\hat{W}_2 \leq W_1$



(b) $W_1 < \hat{W}_2 \leq W_1 + \Theta$



(c) $\hat{W}_2 > W_1 + \Theta$

Figure 1: Illustration of relations between W, W_1, $\hat{W}_2$ and Θ

This relation can be rewritten as:

$$W = \max\left(W_1, \min\left(\hat{W}_2, W_1 + \Theta\right)\right) \ .$$

As a result, we find that

$$\begin{aligned}
\Pr\left[W > w\right] =& \Pr\left[\max\left(W_1, \min\left(\hat{W}_2, W_1 + \Theta\right)\right) > w\right] \\
=& \Pr\left[W_1 > w \vee \min\left(\hat{W}_2, W_1 + \Theta\right) > w\right] \\
=& \Pr\left[W_1 > w\right] + \Pr\left[\min\left(\hat{W}_2, W_1 + \Theta\right) > w\right] \\
& - \Pr\left[W_1 > w \wedge \min\left(\hat{W}_2, W_1 + \Theta\right) > w\right] \\
=& \Pr\left[W_1 > w\right] + \Pr\left[\hat{W}_2 > w \wedge W_1 + \Theta > w\right] \\
& - \Pr\left[W_1 > w \wedge \hat{W}_2 > w \wedge W_1 + \Theta > w\right] \\
=& \Pr\left[W_1 > w\right] + \Pr\left[W_1 + \Theta > w \wedge \hat{W}_2 > w\right] - \Pr\left[W_1 > w \wedge \hat{W}_2 > w\right] \ .
\end{aligned}$$

As calculation of the joint probabilities is difficult, we resort to an approximation: we assume that $\hat{W}_2$ is independent of W_1 and thus of $W_1 + \Theta$ (because W_1 and $\hat{W}_2$ are independent of Θ), leading to the following expression:

$$\Pr\left[W > w\right] \approx \Pr\left[W_1 > w\right] + \Pr\left[\hat{W}_2 > w\right] \left\{\Pr\left[W_1 + \Theta > w\right] - \Pr\left[W_1 > w\right]\right\} \ . \qquad (1)$$

We now calculate $\Pr\left[W_1 > w\right]$, $\Pr\left[W_1 + \Theta > w\right]$ and $\Pr\left[\hat{W}_2 > w\right]$ separately.

3.1. Calculation of $\Pr\left[W_1 > w\right]$

In [22] we have deduced an approximation for $\Pr\left[W_1 > w\right]$ by applying a dominant singularity approximation technique (see e.g. [23]) on the PGF $W_1(z)$ corresponding to W_1. This PGF $W_1(z)$ has been computed in [19] for the simpler model from [22]. Although we include a timer and batch-size dependent service times in this paper, the calculation of $W_1(z)$ corresponding to the current model runs completely along the same lines as in [19]. For this reason, we immediately mention the final expression for $W_1(z)$:

$$W_1(z) = \sum_{k=0}^{c-1} \frac{G_k(z)}{z - A\left(T_c(z)^{1/c}\varepsilon_k\right)} \ ,$$

with ε_k the k-th complex c-th root of 1, i.e.,

$$\varepsilon_k \triangleq e^{\imath 2\pi k/c} \ , \quad 0 \leq k \leq c-1 \ ,$$

and

$$G_k(z) = \frac{T_c(z)-1}{c\lambda T_c(z)} \frac{A\left(T_c(z)^{1/c}\varepsilon_k\right)-1}{\left(T_c(z)^{1/c}\varepsilon_k-1\right)^2} T_c(z)^{1/c}\varepsilon_k \left\{ (z-1)(1-\beta)\sum_{n=0}^{l-1} d(n)\left(T_c(z)^{1/c}\varepsilon_k\right)^n \right.$$
$$\left. +\beta\sum_{n=0}^{l-1} d(n)\left[T_n(z)-\left(T_c(z)^{1/c}\varepsilon_k\right)^n\right] + \sum_{n=l}^{c-1} d(n)\left[T_n(z)-\left(T_c(z)^{1/c}\varepsilon_k\right)^n\right] \right\} \ .$$

$x^{1/c}$ for some $x \in \mathbb{C}$ represents hereby the principal branch of the complex c-th root function, i.e. $x^{1/c} \triangleq |x|^{1/c} e^{\imath \mathrm{Arg}(x)/c}$ with $\mathrm{Arg}(x)$ the principal value of the argument of x, i.e., a mapping in the interval $]-\pi, \pi]$. The probabilities $d(n)$ have to be calculated by solving the set of c linear equations

$$[1-A(z_i)]\sum_{n=0}^{l-1} d(n) z_i^n + \beta\sum_{n=0}^{l-1} d(n)\left[A(z_i)z_i^n - T_n(A(z_i))\right]$$
$$+\sum_{n=l}^{c-1} d(n)[z_i^n - T_n(A(z_i))] = 0 \ , \qquad 1 \le i \le c-1 \ , \tag{2}$$

$$-c + \mathrm{E}\left[T_c\right]\lambda = -c\sum_{n=0}^{l-1} d(n) + \beta\sum_{n=0}^{l-1} d(n)[c + n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]]$$
$$+\sum_{n=l}^{c-1} d(n)[n\mathrm{E}\left[T_c\right] - c\mathrm{E}\left[T_n\right]] \ , \tag{3}$$

with z_i the $c-1$ zeroes of $z^c - T_c(A(z))$ in the open complex unit disk $\{z \in \mathbb{C} : |z| < 1\}$. These zeroes can be calculated one-by-one: each zero z_i is the unique root in the open complex unit disk of the equation

$$z_i = T_c(A(z_i))^{1/c}\varepsilon_i \ .$$

Each of the $c-1$ equations can then be solved by means of a standard root-finding algorithm such as Newton-Raphson. The method also works when $T_c(A(z))$ has infinitely many terms, because the complex unit disk falls within the region of convergence of $T_c(A(z))$.

Remark 1. *The appearance of β in $W_1(z)$ might lead to the premature conjecture that W_1 and Θ are dependent after all. However, although the queueing delay W_1 of the current tagged customer can be influenced by previous timers, the length of the current timer is by no means influenced by W_1: it starts after W_1 with a (geometric) distribution which is independent of W_1.*

In order to apply a dominant singularity approximation, it is necessary to locate the dominant singularities (i.e., those with smallest modulus) of $W_1(z)$ first. As one can observe from the expression for $W_1(z)$, the singularities of $W_1(z)$ might consist of zeroes of $T_c(z)^{1/c}\varepsilon_i - 1$ outside the closed complex unit disk, zeroes of $z - A\left(T_c(z)^{1/c}\varepsilon_i\right)$ outside the closed complex unit disk, possible singularities of $A(T_c(z)^{1/c}\varepsilon_i)$, and possible singularities of $T_n(z)$, for $n = 0, \ldots, c$. We now establish several lemmas that play a crucial role in locating the dominant singularities.

Lemma 1. *The factors* $(T_c(z)^{1/c}\varepsilon_i - 1)^2$, $i = 0, \ldots, c-1$, *in the denominator produce no poles for* $W_1(z)$.

Proof. The proof of this lemma is analogous as for theorem 1 in [22]. □

Lemma 2. *Assumptions 1-5 imply that*
(i) $z^c - T_c(A(z))$ *has exactly one zero in the interval* $]1, \Re[$, *where* $\Re$ *was defined in section 2 as* $\min\{\Re_n : 0 \leq n \leq c\}$, *with* $\Re_n$ *the radius of convergence of* $T_n(A(z))$;
(ii) This zero has multiplicity one;
(iii) $z^c - T_c(A(z))$ *contains no other zeroes with a modulus larger than one and smaller than or equal to this real zero.*

Proof. This lemma has been proved in [24]. □

Let us denote the only zero of $z^c - T_c(A(z))$ in the interval $]1, \Re[$ by $\tilde{z}$. On account of lemma 2 and assumption 3, we have $\tilde{z} < \Re \leq \Re_A$. Hence, the following definition makes sense:

Definition 1.

$$\hat{z} \triangleq A(\tilde{z}) \ .$$

Some useful properties of $\hat{z}$ are mentioned in the following lemma:

Lemma 3. *(i)* $\hat{z} \in \mathbb{R}$;
(ii) $1 < \hat{z} < \Re_T \leq \Re_{T_c}$, *whereby* $\Re_T$ *was defined in section 2 as* $\min\{\Re_{T_n} : 0 \leq n \leq c\}$, *with* $\Re_{T_n}$ *the radius of convergence of* $T_n(z)$.

Proof. (i) Follows naturally from $A(z)$ being a real-valued function within $[1, \Re_A[$ and $\tilde{z}$ being a real number;
(ii) Ensues from $A(1) = 1$, $\tilde{z} > 1$, the PGF $A(z)$ being a monotonically increasing function within $[1, \Re_A[$, and $\tilde{z} < \Re \leq \Re_c$. □

Lemma 4. *Assumptions 1-5 imply that*
(i) $T_c(\hat{z})^{1/c} < \Re_A$ *and* $\hat{z}$ *is a zero of* $z - A(T_c(z)^{1/c})$*;*
(ii) Equations $z - A(T_c(z)^{1/c}\varepsilon_i)$, $i = 0, \ldots, c-1$, *contain no other zeroes with a modulus larger than one and smaller than or equal to* $\hat{z}$*;*
(iii) $\hat{z}$ *is a zero of multiplicity one.*

Proof. The proof of this lemma is analogous as for theorem 2 in [22]. □

Remark 2. *At a cursory glance, one would expect that assumption 5 could be relaxed to:*

$$\lim_{z \uparrow \Re_c} T_c(A(z))/z^c > 1 \ .$$

However, the dominant singularities of $W_1(z)$ *could then stem from singularities of* $T_n(z)$ *for some* n *between* 0 *and* $c-1$ *(in such a case, it would hold that* $R_n < \tilde{z} < R_c$ *and, consequently, because* $\tilde{z} < R_A$, $R_T = R_{T_n} \leq \hat{z} < R_{T_c}$*). Although we thus exclude some PGFs* $T_c(A(z))$*, the commonly adopted PGFs satisfy this assumption. The main advantage is that we can present a general solution whereas otherwise an ad hoc approach would have to be adopted for each PGF* $T_c(A(z))$*.*

This brings us to the following lemma:

Lemma 5. *The dominant singularities of* $W_1(z)$ *do not stem from possible singularities of* $T_n(z)$*.*

Proof. Results from $\hat{z} < \Re_T$ and $\hat{z}$ being a pole of $W_1(z)$. □

Summarizing the lemmas, we obtain the following theorem about the location of the dominant singularities:

Theorem 1. $W_1(z)$ *has one dominant singularity, being a pole* $\hat{z}$*. This dominant pole is a real number larger than one. It is a zero of* $z - A(T_c(z)^{1/c})$*, has multiplicity one, and is equal to* $A(\tilde{z})$*, with* $\tilde{z}$ *the only zero in* $]1, \Re[$ *of* $z^c - T_c(A(z))$*.*

Proof. Ensues directly from lemmas 1-5. □

The zero $\tilde{z}$ can be calculated by means of a standard root-finding algorithm, such as Newton-Raphson, or, because $\tilde{z}$ lies on the real axis, via the bisection method. As $\tilde{z}$ falls within the region of convergence of $T_c(A(z))$, this also works when $T_c(A(z))$ has infinitely many terms.

Now that we have located the dominant singularity of $W_1(z)$, we can apply a dominant singularity approximation (see e.g. [23]). As such, we find the following approximation for $\Pr[W_1 > w]$:

$$\Pr[W_1 > w] \approx \frac{\hat{z}^{-(w+1)}}{1-\hat{z}} \frac{cG_0(\hat{z})}{c - A'\left(T_c(\hat{z})^{1/c}\right) T_c(\hat{z})^{\frac{1}{c}-1} T_c'(\hat{z})} \quad . \tag{4}$$

We have hereby taken into account that $\varepsilon_0 = 1$ by definition.

3.2. Calculation of $\Pr[W_1 + \Theta > w]$

It seems evident to calculate $\Pr[W_1 + \Theta > w]$ by relating it to $\Pr[W_1 > w]$ and taking into account that W_1 and Θ are independent:

$$\Pr[W_1 + \Theta > w] = \sum_{t=0}^{w} \Pr[\Theta = t] \Pr[W_1 > w - t] + \sum_{t=w+1}^{\infty} \Pr[\Theta = t] \quad .$$

In this expression we could then use approximation (4) for $\Pr[W_1 > w]$. *This approach however, might lead to inaccurate results* as for t approaching w, formula (4) for $\Pr[W_1 > w - t]$ can be inaccurate because $w - t$ is very small (dominant singularity approximations for $\Pr[X > w]$ with X some random variable can be inaccurate for small values of w). Let us therefore *consider the PGF corresponding to* $W_1 + \Theta$, which is, because Θ is independent of W_1, equal to $W_1(z)\Theta(z)$, with

$$\Theta(z) \triangleq \mathrm{E}\left[z^{\Theta}\right] = \sum_{n=0}^{\infty} (1-\beta)^n \beta z^n = \frac{\beta}{1-(1-\beta)z} \quad .$$

Note that the dominant singularity of $\Theta(z)$, say z^*, is equal to $1/(1-\beta)$. The dominant singularity of $W_1(z)\Theta(z)$ is thus either equal to z^* (stemming from $\Theta(z)$) or $\hat{z}$ (stemming from $W_1(z)$), depending on which is the smallest. We thus have to consider three scenarios.

$z^* < \hat{z}$

The dominant pole equals z^* and it has multiplicity one. By applying a dominant singularity approximation (see e.g. [23]) we find:

$$\Pr[W_1 + \Theta > w] \approx \frac{(z^*)^{-(w+1)}}{1-z^*} \frac{-\beta}{1-\beta} W_1(z^*) = (1-\beta)^{w+1} W_1\left(\frac{1}{1-\beta}\right) \quad .$$

$z^* > \hat{z}$

In this case, $\hat{z}$ is the dominant pole, with multiplicity one. The dominant singularity approximation leads to

$$\Pr[W_1 + \Theta > w] \approx \frac{\hat{z}^{-(w+1)}}{1-\hat{z}} \frac{\beta}{1-(1-\beta)\hat{z}} \frac{cG_0(\hat{z})}{c - A'\left(T_c(\hat{z})^{1/c}\right) T_c(\hat{z})^{\frac{1}{c}-1} T_c'(\hat{z})} \ .$$

$z^* = \hat{z}$
In this case, $\hat{z} = z^*$ is the dominant pole and it has multiplicity two. The dominant singularity approximation leads in this case to:

$$\Pr[W_1 + \Theta > w] \approx \frac{(z^*)^{-w}}{z^* - 1} G_0(z^*) \beta \frac{-(1-\beta)c}{c - A'(T_c(z^*)^{1/c}) T_c(z^*)^{1/c-1} T_c'(z^*)} w \ . \tag{5}$$

Remark 3. *When $z^* \neq \hat{z}$, it is better to take into account both contributions of z^* and $\hat{z}$ in $\Pr[W_1 > w]$, especially when $z^* \approx \hat{z}$ (see e.g. [25]), leading to:*

$$\begin{aligned} \Pr[W_1 + \Theta > w] \approx & (1-\beta)^{w+1} W_1\left(\frac{1}{1-\beta}\right) \\ & + \frac{\hat{z}^{-(w+1)}}{1-\hat{z}} \frac{\beta}{1-(1-\beta)\hat{z}} \frac{cG_0(\hat{z})}{c - A'\left(T_c(\hat{z})^{1/c}\right) T_c(\hat{z})^{\frac{1}{c}-1} T_c'(\hat{z})} \ . \end{aligned} \tag{6}$$

We adopt this approach in the numerical examples in section 4.

3.3. Calculation of $\Pr\left[\hat{W}_2 > w\right]$

The calculation of $\Pr\left[\hat{W}_2 > w\right]$ runs along the same lines as in [22], leading to

$$\Pr\left[\hat{W}_2 > w\right] = \frac{1}{c} \sum_{m=0}^{l-2} \frac{l-1-m}{m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} A(x)^w g(x) \Big|_{x=0} \ , \tag{7}$$

with

$$\begin{aligned} g(x) &= \frac{1-A(x)}{\lambda(1-x)} \\ &+ \sum_{i=1}^{c-1} \frac{A(\varepsilon_i) - A(x)}{\lambda(\varepsilon_i - x)} \frac{\varepsilon_i(x-1)}{x-\varepsilon_i} \frac{\beta \sum_{n=0}^{l-1} d(n)(\varepsilon_i^n - 1) + \sum_{n=l}^{c-1} d(n)(\varepsilon_i^n - 1)}{A(\varepsilon_i) - 1} \ . \end{aligned}$$

Formula (7) can be implemented in a mathematical program such as matlab. This procedure suffers from the drawback that high-order derivatives may have to be computed, which causes a considerable reduction in speed and even is infeasible if l and c are quite large.

Therefore, we deduce an approximation for $\Pr\left[\hat{W}_2 > w\right]$ whereby no derivatives have to be calculated. Multiplying both sides of (7) by z^w, taking the sum over all values of w, applying Leibniz's rule for the derivative of a product, and taking into account that $[\hat{W}_2(z)-1]/(z-1) = \sum_{w=0}^{\infty} \Pr\left[\hat{W}_2 > w\right] z^w$, we find the following expression for the PGF $\hat{W}_2(z)$ corresponding to $\hat{W}_2$:

$$\frac{\hat{W}_2(z)-1}{z-1} = \frac{1}{c} \sum_{m=0}^{l-2} \frac{l-1-m}{m!} \sum_{j=0}^{m} \frac{C_{m,j}(z)}{[1-zA(0)]^{j+1}} \ , \tag{8}$$

whereby $C_{m,j}(z)$ are functions of z that have no factor $1-zA(0)$ in the denominator. As opposed to $C_{m,j}(z)$ for $j \neq m$, $C_{m,m}(z)$ is relatively easy to calculate:

$$C_{m,m}(z) = m! g(0) z^m A^{'}(0)^m \ .$$

From equation (8), it is clear that $z = 1/A(0)$ is the dominant pole of $[\hat{W}_2(z)-1]/(z-1)$ and that it has multiplicity $l-1$. Analogously as in [22], we retain in (8) for every m the term that produces the largest power of $1-zA(0)$ in the denominator. We thus take advantage of the fact that we can easily calculate $C_{m,m}(z)$ for all m. Hence, in the neighborhood of $z = 1/A(0)$, $[\hat{W}_2(z)-1]/(z-1)$ is proportional to

$$\frac{\hat{W}_2(z)-1}{z-1} \sim \frac{1}{c} \sum_{m=0}^{l-2} \frac{(l-1-m) g(0) z^m A^{'}(0)^m}{[1-zA(0)]^{m+1}} \ . \tag{9}$$

Next, as

$$\frac{1}{[1-zA(0)]^{m+1}} = \frac{1}{m! A(0)^m} \sum_{w=m}^{\infty} A(0)^w z^{w-m} \frac{w!}{(w-m)!} \ ,$$

(9) transforms into:

$$\begin{aligned} \frac{\hat{W}_2(z)-1}{z-1} &\sim \frac{g(0)}{c} \sum_{m=0}^{l-2} A^{'}(0)^m (l-1-m) \sum_{w=m}^{\infty} z^w \frac{w!}{m!(w-m)!} A(0)^{w-m} \\ &= \frac{g(0)}{c} \sum_{w=0}^{\infty} z^w \sum_{m=0}^{\min(l-2,w)} A^{'}(0)^m (l-1-m) \binom{w}{m} A(0)^{w-m} \ . \end{aligned}$$

Equating powers of z^w at both sides of the equation finally yields

$$\Pr\left[\hat{W}_2 > w\right] \approx \frac{g(0)}{c} \sum_{m=0}^{\min(l-2,w)} A^{'}(0)^m (l-1-m) \binom{w}{m} A(0)^{w-m} \ . \tag{10}$$

Note that for large w, formula (10) becomes a sum from 0 to $l-2$. We further point out that the binomial coefficient causes no difficulties, since efficient routines exist to calculate them, even for large w.

Remark 4. *As in [22], this approach is not suited for cases whereby* $A^{'}(0) = 0$*, as only the term corresponding to* $m = 0$ *in (9) differs from 0. In these cases, additional terms with* $j < m$ *must be taken into account from (8).*

3.4. Summary of computation steps

To close this section, we give a brief overview of the steps required to calculate the approximation for $\Pr[W > w]$:

- Calculate the $c-1$ zeroes $z_1, \ldots, z_{c-1}$ of $z^c - T_c(A(z))$ inside the open complex unit disk $\{z \in \mathbb{C} : |z| < 1\}$;
- Solve the set of c linear equations (2)-(3) in the c unknown probabilities $d(0), \ldots d(c-1)$;
- Calculate $\tilde{z}$: the unique zero in $]1, \Re[$ of $z^c - T_c(A(z))$ and set $\hat{z} = A(\tilde{z})$;
- Calculate $\Pr[W_1 > w]$ via formula (4);
- Calculate $\Pr[W_1 + \Theta > w]$:
  - If $z^* \triangleq 1/(1-\beta) = \hat{z}$: use formula (5);
  - Else: use formula (6);
- Calculate $\Pr\left[\hat{W}_2 > w\right]$ via formula (10);
- Calculate $\Pr[W > w]$ via formula (1) and invoke the calculated values for $\Pr[W_1 > w]$, $\Pr[W_1 + \Theta > w]$ and $\Pr\left[\hat{W}_2 > w\right]$.

4. Evaluation of approximation formulas

In this section, we evaluate the accuracy of our approach. We have therefore studied an extensive set of examples, corresponding to several values of l, c and β, and various combinations of distributions for the number of customer arrivals ("Poisson" $A(z) = e^{\lambda(z-1)}$; "Geometric" $A(z) = 1/(1+\lambda-\lambda z)$; "C-center" $A(z) = 1 - \lambda/c + \lambda/(2c)(z^{c-1} + z^{c+1})$) and service times ("Geometric" $T_n(z) = z/[\mathrm{E}[T_n] + (1 - \mathrm{E}[T_n])z]$; "25" $T_n(z) = (25 - \mathrm{E}[T_c])z/24 +$

$(\mathrm{E}[T_c]-1)z^{25}/24)$. We believe that these combinations cover both commonly adopted (Poisson, geometric) as well as more special (burstier) distributions. We prefer not to present all these examples, in order to keep the paper concise. Instead, we summarize our findings and we focus on the circumstances where one should be careful. The approximations and the real values of $\Pr[W_1 > w]$, $\Pr[W_1 + \Theta > w]$, $\Pr\left[\hat{W}_2 > w\right]$ and $\Pr[W > w]$ are depicted versus the load in Figures 2-7 for the peculiar situations.
Note that only for $\Pr\left[\hat{W}_2 > w\right]$ we have an exact formula at our disposal (formula (7)). In order to evaluate the other tail probabilities, we show the mean values resulting from 20 Monte Carlo simulations whereby each simulation generates W_1, $W_1+\Theta$ and W for 10^9 customers (we do not plot the entire confidence intervals to enhance the readibility of the figures and because the confidence intervals are very small anyway). We discuss successively the accuracy of formula (4) for $\Pr[W_1 > w]$ (section 4.1), expressions (5)-(6) for $\Pr[W_1 + \Theta > w]$ (section 4.2), approximation (10) for $\Pr\left[\hat{W}_2 > w\right]$ (section 4.3) and expression (1) for $\Pr[W > w]$ (section 4.4).

4.1. $\Pr[W_1 > w]$

Approximation (4) is very accurate. Only when the load (and thus the mean arrival rate) is small, the approximation can become inaccurate. We observe from Figures 2-7 that the approximation is very accurate in case of service times of 1 or 25 slots, whereas it is not always accurate in case of geometric services. The key issue is that in the latter case, each $T_n(z)$ has a singularity $\gamma_n \triangleq \mathrm{E}[T_n]/[\mathrm{E}[T_n]-1]$. In order to gain deeper insight, we have reported in Table 1 the dominant pole $\hat{z}$ of $W_1(z)$ versus the load, and γ_n versus n. We notice that, in case of Poisson and geometric arrivals, the smaller the load, the more $\hat{z}$ approaches to the singularity γ_c. This is the reason why the approximation becomes inaccurate for small loads in these cases. This anomaly is not specific for our model but is inherent to approximations based on dominant singularities in general. In case of c-centered arrivals, $\hat{z}$ approaches γ_n a lot slower, which entails a much better accuracy of the approximation (see Fig. 5). Fig. 3 also exhibits that the approximation is precise in case of Poisson arrivals and services of either 1 or 25 slots for all values of the load. The reason is that $T_n(z)$ has no singularities in this case, as explained above.
In general, the *approximation is accurate except when the load is small in*

combination with $T_n(z)$ (and/or $A(z)$) having singularities. In such situations, it is possible to enhance the approximation by adopting an ad hoc approach whereby the contributions of the other singularity(ies) nearby $\hat{z}$ is (are) also incorporated (see e.g. [25]).

Remark 5. *In [22], we have also considered an example with Poisson arrivals and geometric service times. There however, the approximation for* $\Pr[W_1 > w]$ *did not suffer as much as here from the singularity of the PGF of the service times. The reason for this is that here we deal with several PGFs of the service times depending on the number of customers in the served batch, and that the dominant pole* $\hat{z}$ *of* $W_1(z)$ *also approaches the singularities of* $T_{c-1}(z)$, $T_{c-2}(z)$, *et cetera when the load becomes very small. It has been shown that approximations based on dominant singularities become less accurate the more other singularities approach the dominant singularity(ies) (see e.g. [25]).*

4.2. $\Pr[W_1 + \Theta > w]$

Approximations (5)-(6) for $\Pr[W_1 + \Theta > w]$ are accurate in all scenarios. Indeed, for larger loads, W_1 is dominant in $W_1 + \Theta$ and, as discussed before, the approximation for $\Pr[W_1 > w]$ is precise for larger loads, whereas for smaller loads, W_1 becomes small, so that Θ determines the behaviour of $W_1 + \Theta$ and the formula for $\Pr[\Theta > w]$ is exact (due to its geometric distribution).

In some special cases however, the approximation might become inaccurate for smaller loads. Consider for instance the system with Poisson arrivals, geometric service times with $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$ and $\beta = 0.2$. In this situation, the singularity of $\Theta(z)$ equals $1/(1-\beta) = 1.25$ and the singularity γ_c of $T_c(z)$ equals $1.11\ldots$ (see Table 1). Hence, $1/(1-\beta)$ is, regardless of the load, larger than $\hat{z}$ (because $\hat{z} < \gamma_c$), which means that $\hat{z}$ is always the dominant pole of $W_1(z)\Theta(z)$. Hence, W_1 even dominates $W_1 + \Theta$ for smaller loads, which results in an inaccurate approximation (see Fig. 6). When, on the other hand, $\mathrm{E}[T_n]$ is equal to $3 + 0.1n$, γ_c is equal to $1.333\ldots$, so that Θ will again dominate for smaller loads, which leads to a good approximation (see Fig. 7).

Hence, *approximations (5)-(6) for* $\Pr[W_1 + \Theta > w]$ *are accurate, except for special cases whereby* $1/(1-\beta)$ *is always larger than* $\hat{z}$, *which can lead to awkward results for small loads.* Again, this can be resolved by following an ad hoc approach in such situations.

4.3. $\Pr\left[\hat{W}_2 > w\right]$

Approximation (10) is very accurate for larger values of w, except in the case of c-centered arrivals (see Fig. 5). The reason is that $A^{'}(0) = 0$ in that case, and we have explained in remark 4 that approximation formula (10) can become inaccurate in such a case. We can thus conclude that *approximation (10) is extremely suited for quickly assessing the order of magnitude of* $\Pr\left[\hat{W}_2 > w\right]$*, except when* $A^{'}(0) = 0$.

4.4. $\Pr[W > w]$

Finally, we discuss the accuracy of approximation (1) for $\Pr[W > w]$. *Approximation (1) is accurate, except for values of the load between, roughly speaking, 0.15 and 0.35, where it is less precise but still acceptable for the purpose of assessing the order of magnitude of* $\Pr[W > w]$. This can be observed from Figures 2-7 (we have utilized the approximations for $\Pr[W_1 > w]$, $\Pr[W_1 + \Theta > w]$ and $\Pr\left[\hat{W}_2 > w\right]$ in formula (1)).
The approximation being extremely accurate for larger loads follows from $\Pr[W_1 > w]$ then dominating in (1) and the approximation of $\Pr[W_1 > w]$ being outstanding in this area. For "medium" values of the load, $\Pr[W_1 > w] \approx \Pr\left[\hat{W}_2 > w\right]$, so that $\Pr[W_1 > w]$ still plays a considerable role in (1). As explained in subsection 4.1, the approximation for $\Pr[W_1 > w]$ for geometric service times combined with Poisson or geometric arrivals is not excellent but still adequate in this area. When the load is small, $\Pr[W_1 > w] << \Pr\left[\hat{W}_2 > w\right]$, so that $\Pr\left[\hat{W}_2 > w\right]$ and $\Pr[W_1 + \Theta > w]$ dominate in (1). In addition, as the load is small, it generally holds that $\Pr[W_1 + \Theta > w] \approx \Pr[\Theta > w]$ (except in some special cases, of which we discuss one at the end of this subsection). As the approximation for $\Pr\left[\hat{W}_2 > w\right]$ is precise and the formula for $\Pr[\Theta > w]$ is exact (it has a geometric distribution), the approximation is very accurate.

Let us now study the accuracy in the very special case of c-centered arrivals (where either 0, $c-1$ or $c+1$ customers arrive in a slot) in combination with $l > 1$ and $\beta \neq 0$ (see Figure 5). We observe that, although we adopt

exact expression (7) for $\Pr\left[\hat{W}_2 > w\right]$ in approximation (1) for $\Pr[W > w]$, approximation (1) is inaccurate for smaller values of the load. The reason is that approximation (1) is mainly based on the assumption that $\hat{W}_2$ is independent of W_1, which is not a good assumption in this peculiar example. Indeed, on account of the low load, the system is very likely to be empty at slot mark J (we here denote the slot wherein the tagged customer arrives by J). In such a case, W_1 can only differ from zero when $c+1$ customers arrive during slot J and if the tagged customer is the final arrival in that slot. As a result, the batch wherein the tagged customer will be served, only contains the tagged customer itself at slot mark $J+1$, which means that $\hat{W}_2 > 0$. When, on the other hand, $W_1 = 0$, $\hat{W}_2$ can only differ from zero when $c-1$ customers arrive during slot J and if $l = c$. In other words, W_1 and $\hat{W}_2$ are strongly correlated in this specific case.

Before closing this section, we return to the examples where $\beta = 0.2$. When $\mathrm{E}[T_n] = 8 + 0.2n$ (Fig. 6), the approximation for $\Pr[W > w]$ is inaccurate for small loads, which is a direct result of W_1 dominating over Θ and approximation (4) for $\Pr[W_1 > w]$ being inaccurate in this case. On the other hand, when $\mathrm{E}[T_n] = 3 + 0.1n$ (Fig. 7), Θ again dominates over W_1 for small loads, which results in a good approximation.

5. Importance of the model

As compared to the model of our previous paper [22], we have included batch-size dependent service times and a timer mechanism in this paper. In this section, we briefly demonstrate that ignoring these features can lead to very distinct results, which highlights the importance of the current paper. In Fig. 8, we have depicted $\Pr[W > w]$ versus w and versus ρ, both for $\mathrm{E}[T_n] = 1 + 0.9n$ (i.e., batch-size dependent) and $\mathrm{E}[T_n] = 10$ (i.e., batch-size independent) geometric service times. We have set $c = 10$ so that $\mathrm{E}[T_c] = 10$ in both scenarios. In addition, we have considered Poisson arrivals, $l = 3$, and $\beta = 0.01$. In this example, the average service time of a batch with less than c customers is smaller in case of batch-size dependent service times. We thus expect smaller tail probabilities of the customer delay in that case. Hence, the only remaining question sounds: is the difference significant? We observe from Fig. 8 that in cases of very small and very large load, the difference is small. This is because the timer dominates the delay in case of very small load and the timer is independent of the distribution of the service

times. In case of very large load, the server almost always serves batches of c customers, which leads to the same average service time. However, Fig. 8 exhibits that the batch-size dependent service times do play a significant role for all other loads. For $\rho \approx 0.3$ the difference even mounts to three orders of magnitude. Ignoring batch-size dependent service times can thus lead to very distorted results.
Let us now examine the influence of the timer mechanism. In Fig. 9, $\Pr[W > w]$ is shown versus w and versus ρ both for $\beta = 0$ (i.e., without timer mechanism) and $\beta = 0.2$ (i.e., with timer mechanism) in case of $l = 5$, $c = 10$, Poisson arrivals, and geometric service times with $\mathrm{E}[T_n] = 8 + 0.2n$. We observe that the larger w and the smaller ρ, the more pronounced is the benefit of adopting $\beta = 0.2$ instead of $\beta = 0$. The difference even mounts to many orders of magnitude. This confirms that a timer is very effective to avoid excessive delays due to postponing service in case of light traffic.
We can thus conclude that these examples clearly reflect that *batch-size dependent service times and a timer mechanism can have a major impact on the tail probabilities of the customer delay.* This highlights the importance of the inclusion of these features in the studied model.

6. Conclusions

In this paper, we have deduced approximations for the tail probabilities of the customer delay in a queueing model with batch arrivals and batch service. As compared to our previous paper [22], we have included two very important features in the model. We have considered a general dependency between the service time of a batch and the number of customers within it, and we have incorporated a timer mechanism as well. We have demonstrated that deducing approximations for this extended model entails various pitfalls that would lead to inaccurate approximations and we have dealt with those pitfalls. We have also evaluated the performance of our approach in much more detail in this paper. We feel that it is justified to state that our approximations are very useful for the purpose of assessing the order of magnitude of the customer delay, except in some special cases that have been discussed extensively. Finally, we have illustrated that neglecting batch-size dependent service times or a timer mechanism can lead to a devastating assessment of the tail probabilities of the customer delay, which highlights the necessity to include these features in the model.

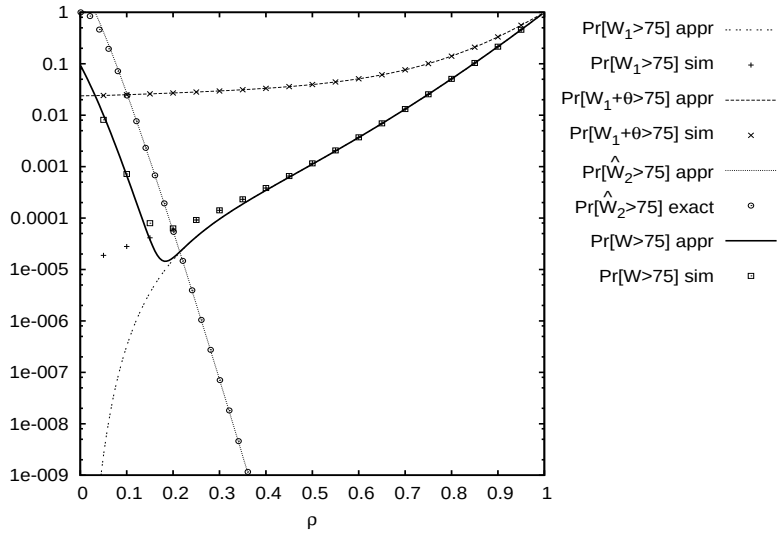


Figure 2: Evaluation of the approximation formulas; Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$, $\beta = 0.05$

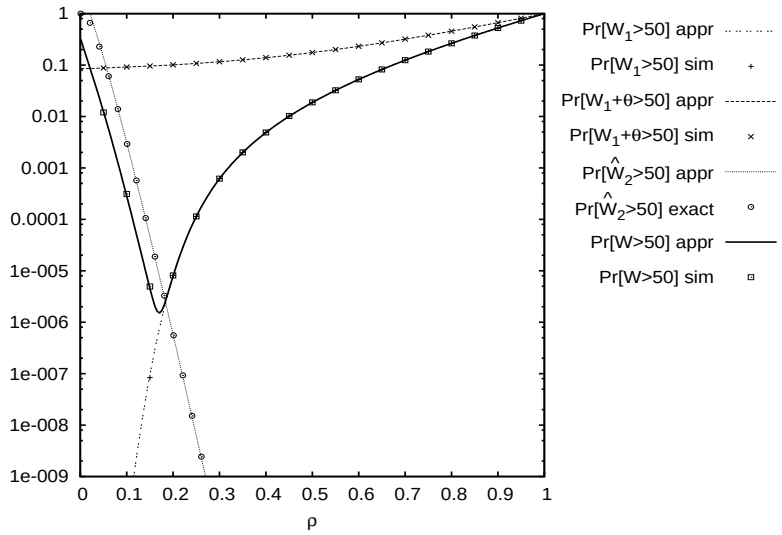


Figure 3: Evaluation of the approximation formulas; Poisson arrivals, 1 or 25 slots service $\mathrm{E}[T_n] = 5$, $c = 10$, $l = 5$, $\beta = 0.05$

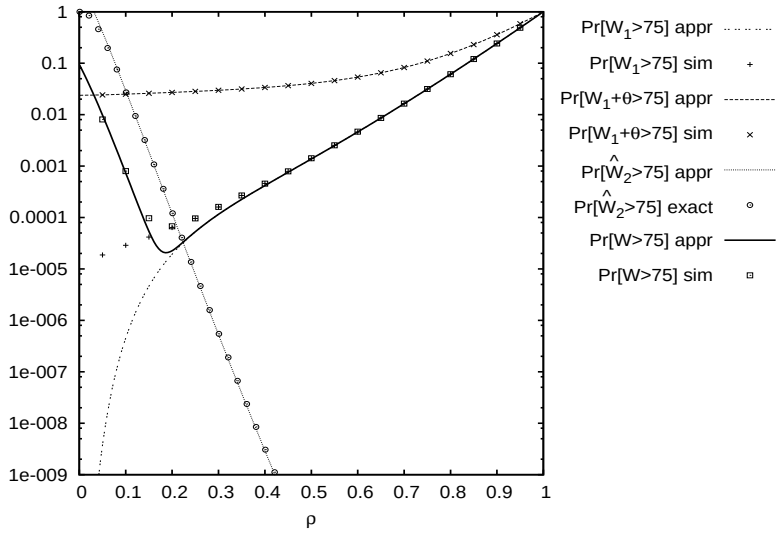


Figure 4: Evaluation of the approximation formulas; geometric arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$, $\beta = 0.05$

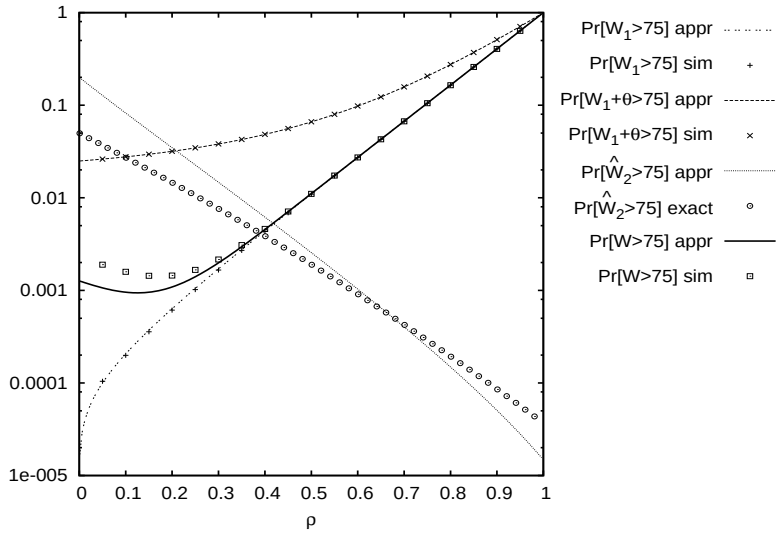


Figure 5: Evaluation of the approximation formulas; c-centered arrivals (exact formula (7) for $\Pr\left[\hat{W}_2 > w\right]$ is used in formula (1)), geometric services $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$, $\beta = 0.05$

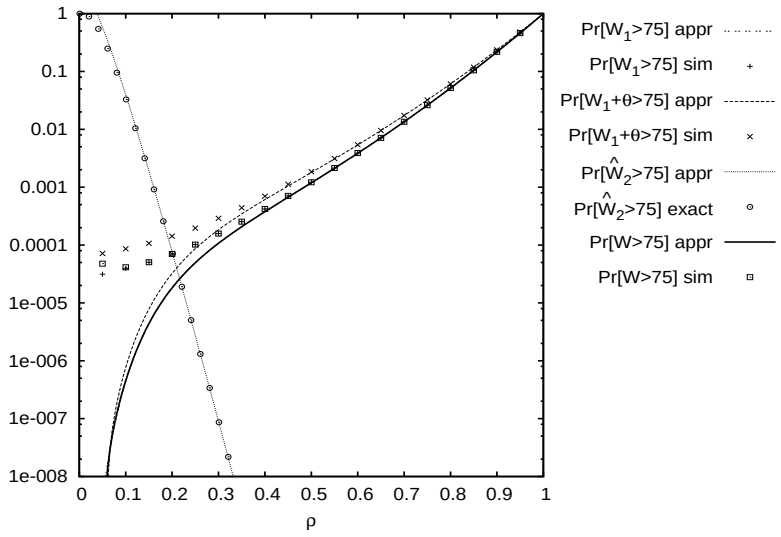


Figure 6: Evaluation of the approximation formulas; Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$, $\beta = 0.2$

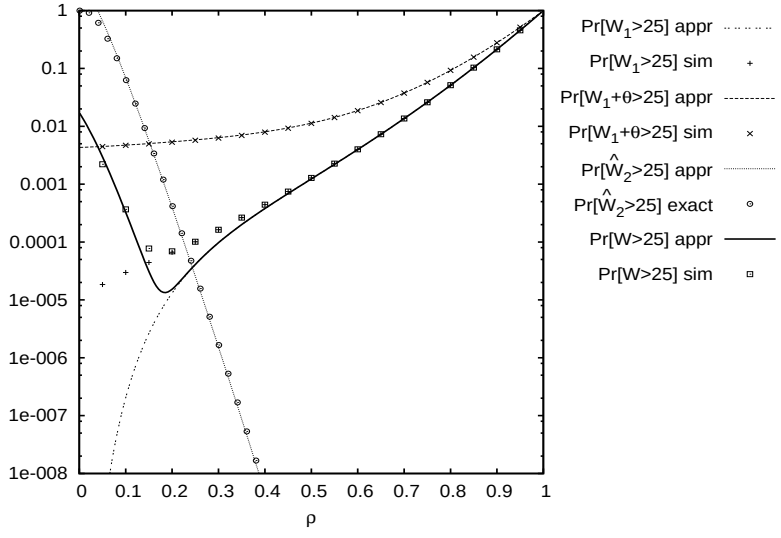
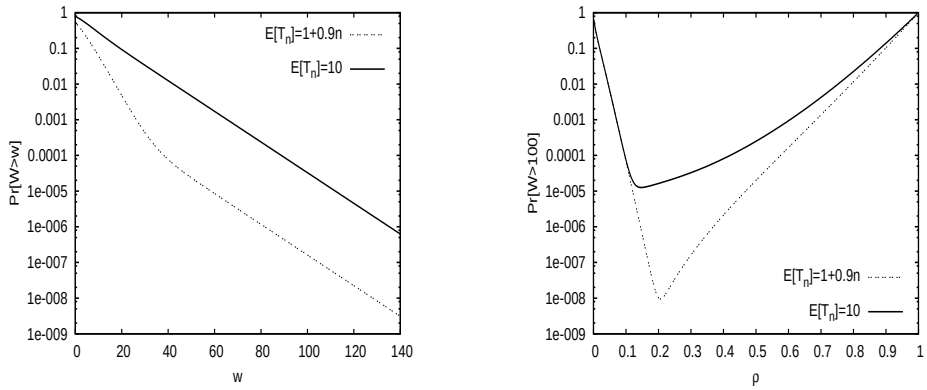


Figure 7: Evaluation of the approximation formulas; Poisson arrivals, geometric services $\mathrm{E}[T_n] = 3 + 0.1n$, $c = 10$, $l = 5$, $\beta = 0.2$

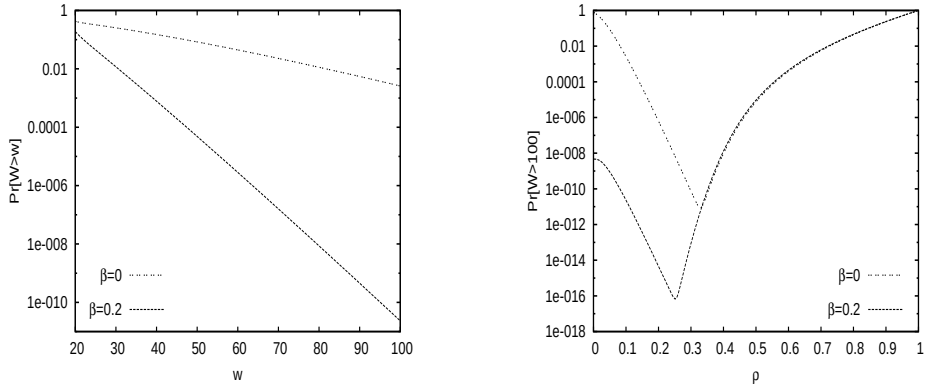
Table 1: Singularities γ_n of $T_n(z)$ versus n, and dominant pole $\hat{z}$ of $W_1(z)$ versus the load ρ, for several distributions of $A(z)$; $c = 10$, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$

n	γ_n	ρ	$\hat{z}$ Poisson	$\hat{z}$ geometric	$\hat{z}$ c-centered
0	1.142857142857	0.9	1.019517053853	1.017984717482	1.011049828292
2	1.135135135135	0.7	1.055046820392	1.052088641246	1.033151838444
4	1.128205128205	0.5	1.084194576530	1.081656490042	1.055263630655
6	1.121951219512	0.3	1.104020768326	1.102948002054	1.077404642357
8	1.116279069767	0.1	1.111018197772	1.110989967865	1.099654929738
9	1.113636363636	0.05	1.111109639324	1.111109020127	1.105282554417
10	1.111111111111	0.01	1.111111111100	1.111111111106	1.109878832698



(a) versus w; $\rho = 0.3$ (b) versus ρ

Figure 8: Influence of batch-size dependent service times; Poisson arrivals, geometric services, $c = 10$, $l = 3$, $\beta = 0.01$



(a) versus w; $\rho = 0.1$ (b) versus ρ

Figure 9: Influence of timer mechanism; Poisson arrivals, geometric services $\mathrm{E}[T_n] = 8 + 0.2n$, $c = 10$, $l = 5$

References

[1] Y. Chen, C. Qiao, X. Yu, Optical burst switching (OBS): a new area in optical networking research, IEEE Network 18(3) (2004) 16–23.

[2] C. Qiao, M. Yoo, Optical burst switching (OBS) - a new paradigm for an optical internet, Journal of High Speed Networks 8 (1) (1999) 69–84.

[3] K. Lu, D. Wu, Y. Fang, R. Qiu, Performance analysis of a burst-frame-based MAC protocol for ultra-wideband ad hoc networks, in: Proceedings of the IEEE International Conference on Communications (ICC 2005), Seoul, May 16-20, Vol.5, 2005, pp. 2937–2941.

[4] B. Bellalta, A queueing model for the non-continuous frame assembly scheme in finite buffers, in: Proceedings of the 16th international conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA 2009), 2009, pp. 219–233.

[5] A. Banerjee, K. Sikdar, U. Gupta, Computing system length distribution of a finite-buffer bulk-arrival bulk-service queue with variable server capacity, International Journal of Operational Research 12(3) (2011) 294–317.

[6] A. Banerjee, U. Gupta, Reducing congestion in bulk-service finite-buffer queueing system using batch-size-dependent service, Performance Evaluation 69(1) (2012) 53–70.

[7] R. Arumuganathan, S. Jeyakumar, Steady state analysis of a bulk queue with multiple vacations, setup times with n-policy and closedown times, Applied Mathematical Modelling 29 (2005) 972–986.

[8] S. Chang, D. Choi, Performance analysis of a finite-buffer discrete-time queue with bulk arrival, bulk service and vacations, Computers and Operations Research 32 (2005) 2213–2234.

[9] S. Chang, T. Takine, Factorization and stochastic decomposition properties in bulk queues with generalized vacations, Queueing Systems 50 (2005) 165–183.

[10] V. Goswami, J. Mohanty, S. Samanta, Discrete-time bulk-service queues with accessible and non-accessible batches, Applied Mathematics and Computation 182 (2006) 898–906.

[11] U. Gupta, V. Goswami, Performance analysis of finite buffer discrete-time queue with bulk service, Computers and Operations Research 29 (2002) 1331–1341.

[12] A. Janssen, J. van Leeuwaarden, Analytic computation schemes for the discrete-time bulk service queue, Queueing Systems 50 (2005) 141–163.

[13] W. Powell, P. Humblet, The bulk service queue with a general control strategy: theoretical analysis and a new computational procedure, Operations Research 34 (2) (1986) 267–275.

[14] S. Samanta, M. Chaudhry, U. Gupta, Discrete-time $Geo^x|G^{(a,b)}|1|n$ queues with single and multiple vacations, Mathematical and Computer Modelling 45 (2007) 93–108.

[15] K. Sikdar, U. Gupta, Analytic and numerical aspects of batch service queues with single vacation, Computers and Operations Research 32 (2005) 943–966.

[16] X. Yi, N. Kim, B. Yoon, K. Chae, Analysis of the queue-length distribution for the discrete-time batch-service $Geo/G^{a,Y}/1/K$ queue, European Journal of Operational Research 181 (2007) 787–792.

[17] D. Claeys, K. Laevens, J. Walraevens, H. Bruneel, Delay in a discrete-time queueing model with batch arrivals and batch services, in: Proceedings of the Information Technology: New Generations Conference (ITNG 2008), 2008, pp. 1040–1045.

[18] D. Claeys, J. Walraevens, K. Laevens, H. Bruneel, Delay analysis of two batch-service queueing models with batch arrivals: $Geo^X/Geo^c/1$, 4OR 8(3) (2010) 255–269.

[19] D. Claeys, J. Walraevens, K. Laevens, H. Bruneel, Analysis of threshold-based batch-service queueing systems with batch arrivals and general service times, Performance Evaluation 68(6) (2011) 528–549.

[20] D. D. Vleeschauwer, A. V. Moffaert, M. Büchli, G. Petit, B. Steyaert, H. Bruneel, Determining the tolerable load generated by a set of packet-based phones on a multiplexing node, in: Proceedings of the 17th International Teletraffic Congress (ITC17), 2001.

[21] D. Claeys, K. Laevens, J. Walraevens, H. Bruneel, Complete characterisation of the customer delay in a queueing system with batch arrivals and batch service, Mathematical Methods of Operations Research 72(1) (2010) 1–23.

[22] D. Claeys, B. Steyaert, J. Walraevens, K. Laevens, H. Bruneel, Tail distribution of the delay in a general batch-service queueing model, Computers and Operations Research 39 (2012) 2733–2741.

[23] H. Bruneel, B. Steyaert, E. Desmet, G. Petit, Analytic derivation of tail probabilities for queue lengths and waiting times in ATM multiserver queues, European Journal of Operational Research 76 (1994) 563–572.

[24] B. Steyaert, Analysis of generic discrete-time buffer models with irregular packet arrival patterns, Ph.D. thesis, Ghent University (2008).

[25] B. Steyaert, J. Walraevens, D. Fiems, D. D. Vleeschauwer, H. Bruneel, Heterogeneous sources model for DSL access multiplexers, Electronics Letters 44(21) (2008) 1282–1283.